



CORPUS LINGUISTICS

G. Richard Bevan
December 12, 2019

Rachel Maddow Faces Slapdown by UC Linguistics Professor in Defamation Suit

POSTED BY KEN STONE ON DECEMBER 2, 2019 IN BUSINESS | 10796 VIEWS | 1 COMMENTS | [LEAVE A COMMENT](#)

Share This Article:



Lawyers for MSNBC host Rachel Maddow (inset) will have to deal with arguments by UC Santa Barbara linguistics professor Stefan Gries (left) that she indeed was taken literally by viewers in labeling One America News as Russian propaganda. Images via uchicago.edu and Twitter

By Ken Stone

If **Rachel Maddow** loses her \$10 million defamation case brought by the owner of San Diego-based **One America News Network**, she can blame, in part, a UC Santa Barbara linguistics professor.

Support Times of San Diego's growth
with a small monthly contribution

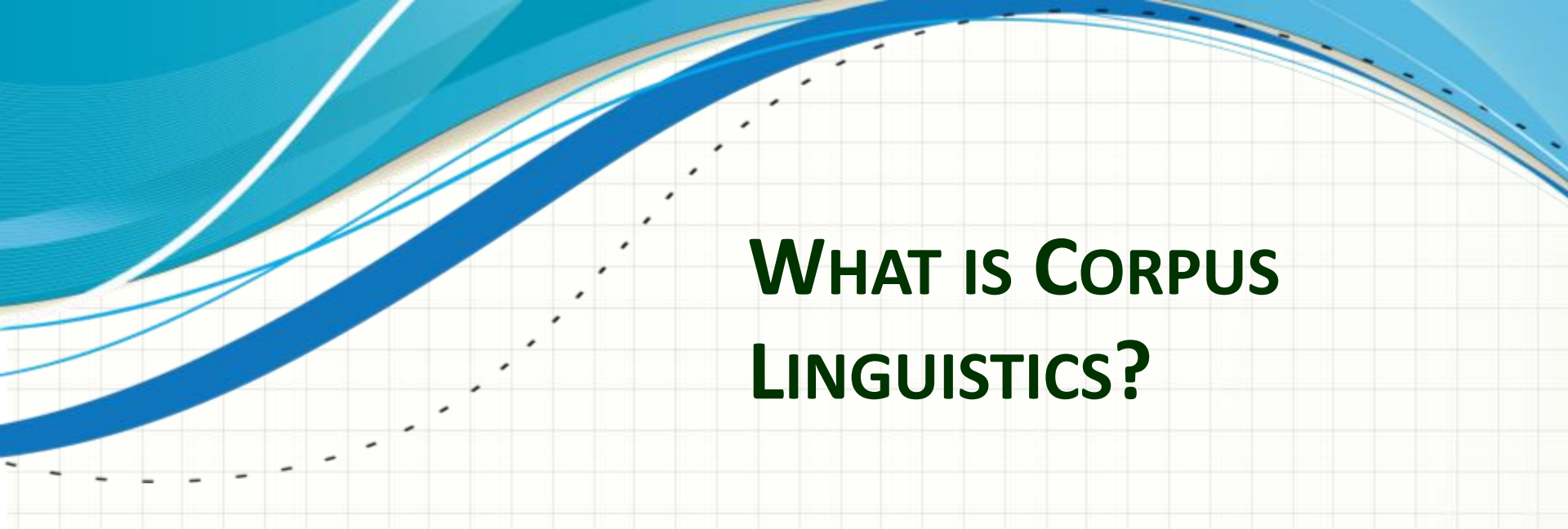
The professor, **Stefan Thomas Gries**, argues in a long analysis of Maddow's on-air speech patterns that when she says "literally," she means "in fact."

YOU KEEP USING THAT WORD

**I DO NOT THINK IT MEANS
WHAT YOU THINK IT MEANS**


Corpus Linguistics Defined

- Linguistics: The scientific study of language and its structure, including the study of morphology, syntax, phonetics, and semantics.
- Corpora: Computerized Databases containing millions or billions of naturally occurring text from various sources.




WHAT IS CORPUS LINGUISTICS?

Corpus linguistics is the analysis of naturally occurring language on the basis of electronic databases known as corpora. The analysis is performed with the help of a computer, with specialized software, and takes into account natural word usage in the context of linguistic usage patterns.



“Linguistic” Questions Asked By Judges

- What is the “ordinary meaning” of a term?
- Is a word ambiguous?
- Has a word or phrase acquired a “specialized meaning within a particular industry?”
- What is the “original public meaning” of a Constitutional phrase?
- Does the agreement have a “plain meaning”?



How do Judges Typically Answer “Linguistic” Questions?

- Intuition
- Dictionaries
- Etymology
(origin of words)
- Morphology
(study of the
forms of words).

The Path Forward in Today's Technological World

Artificial Intelligence Is on the Case in the Legal Profession

By Thomas J. Lane • 10/16/13 9:30am



68,762 views | May 23, 2018, 12:38am

How AI And Machine Learning Are Transforming Law Firms And The Legal Sector



Bernard Marr Contributor @
Enterprise Tech

Whenever a professional sector faces new technology, questions arise regarding how that technology will disrupt daily operations and the careers of those who choose that profession. And lawyers and the legal profession are no exception. Today, artificial intelligence (AI) is beginning to transform the legal profession in many ways, but in most cases it augments what humans do and frees them up to take on higher-level tasks such as advising to clients, negotiating deals and appearing in court.




BYU CORPORA

Corpus (click to use)	Description
iWeb	14 billion words from the Web
COCA	520 million, US, 1990-2015
COHA	400 million, US, 1810s-2000s
NOW	6.0 billion, Web news, 2010-yesterday
GloWbE	1.9 billion, Web, 20 countries
Wikipedia	1.9 billion, Wikipedia
BYU-BNC	100 million, British, 1980s-1993
TIME	100 million, US, 1923-2006
SOAP	100 million, US, 1990s-2000s
Strathy	50 million, Canada
CORE	53 million words, Web genres
US Supreme Court	130 million words, 1800s-2000s

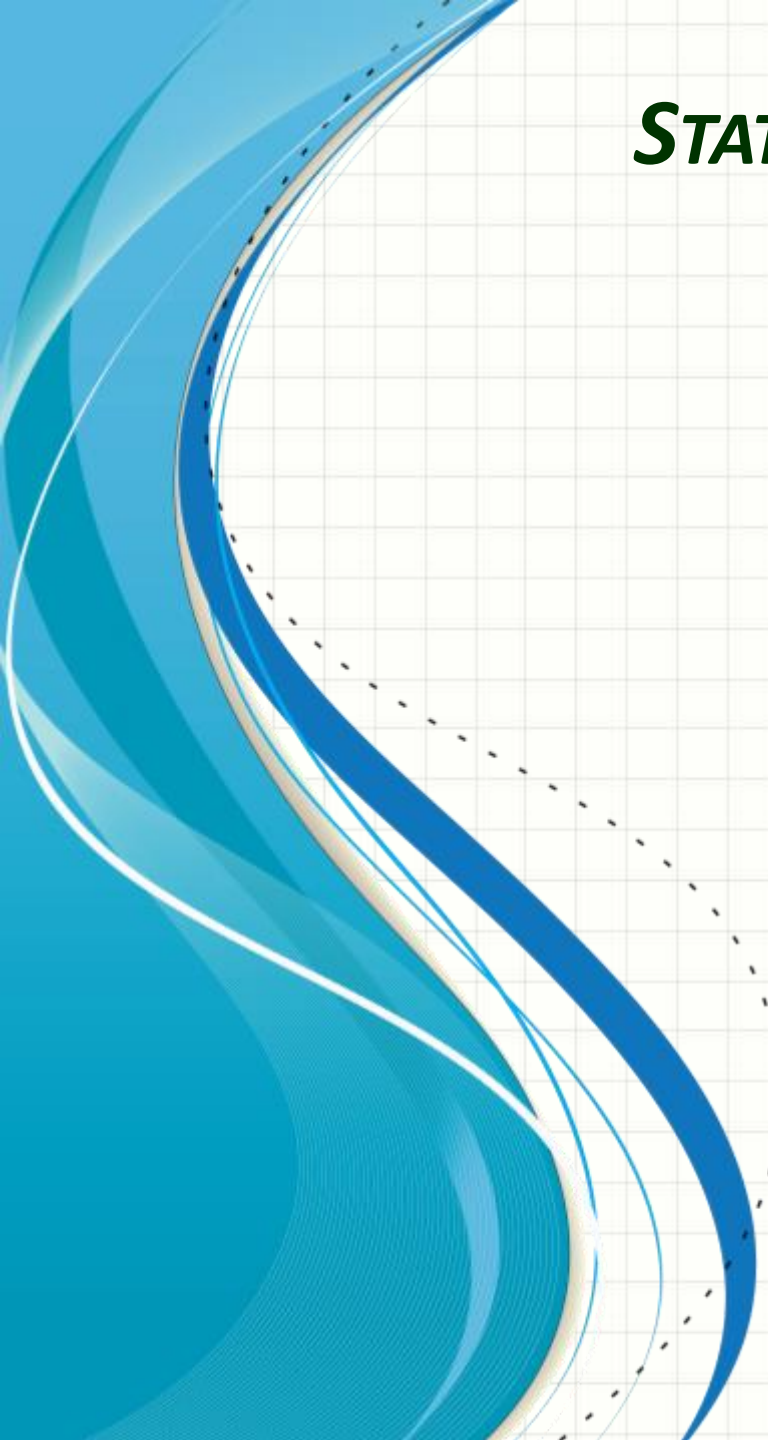
<https://www.english-corpora.org/>



The [most widely used online corpora](#). [Overview](#), [search types](#), [looking at variation](#), [corpus-based resources](#).

The links below are for the online interface. But you can also  download the corpora for use on your own computer.

Corpus (online access)	Download	# words	Dialect	Time period	Genre(s)
iWeb: The Intelligent Web-based Corpus		14 billion	6 countries	2017	Web
News on the Web (NOW)		8.7 billion+	20 countries	2010-last month	Web: News
Global Web-Based English (GloWbE)		1.9 billion	20 countries	2012-13	Web (incl blogs)
Wikipedia Corpus		1.9 billion	(Various)	2014	Wikipedia
Corpus of Contemporary American English (COCA)		560 million	American	1990-2017	Balanced
Corpus of Historical American English (COHA)		400 million	American	1810-2009	Balanced
The TV Corpus		325 million	6 countries	1950-2018	TV shows
The Movie Corpus		200 million	6 countries	1930-2018	Movies
Corpus of American Soap Operas		100 million	American	2001-2012	TV shows
Hansard Corpus		1.6 billion	British	1803-2005	Parliament
Early English Books Online		755 million	British	1470s-1690s	(Various)
Corpus of US Supreme Court Opinions		130 million	American	1790s-present	Legal opinions
TIME Magazine Corpus		100 million	American	1923-2006	Magazine
British National Corpus (BNC) *		100 million	British	1980s-1993	Balanced
Strathy Corpus (Canada)		50 million	Canadian	1970s-2000s	Balanced
CORE Corpus		50 million	6 countries	2014	Web
From Google Books n-grams (compare)					
American English		155 billion	American	1500s-2000s	(Various)



STATE V. LANTIS, 165 IDAHO 427, 447 P.3D 875, 880 (2019)

Corpus linguistic analysis supported the Court's reasoning -- helping to analyze the meaning of "disturb the peace," utilizing COHA, the Corpus of Historical American English, which contains 400 million words which are searchable in context.

List Chart **Collocates** Compare KWIC Word/phrase [POS] Collocates [POS]

+	6	5	4	3	2	1	0	0	1	2	3	4	5	6	+
---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---

☒ Sections Texts/Virtual Sort/Limit **Options**# HITS # KWIC GROUP BY DISPLAY SAVE LISTS 

(HIDE HELP)

NOT LOGGED IN

OTHER OPTIONS

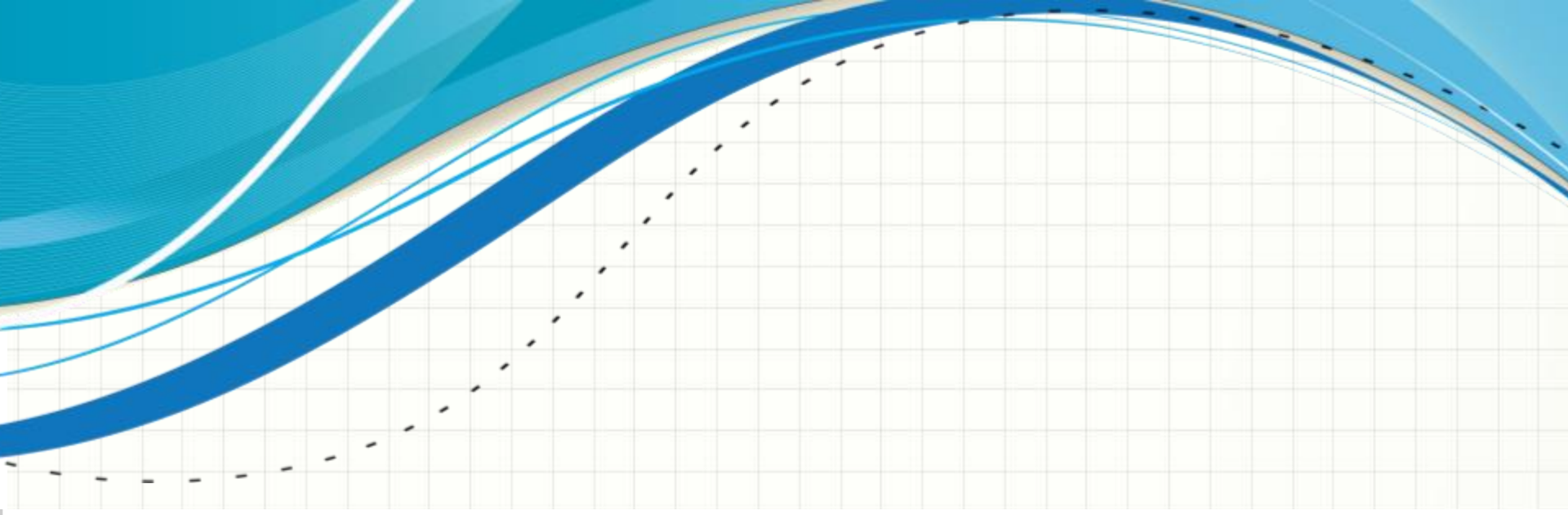
HITS is the number of results.

KWIC is the number of results for a KWIC (concordances) search.

GROUP BY determines whether words are grouped by word form (e.g. *decide* and *decided* separately), lemma (e.g. all forms of *decide* together), and whether you see the part of speech for word (e.g. *beat* as a noun and verb displayed separately).

DISPLAY shows raw frequency, occurrences per million words, or a combination of these.

SAVE LISTS allows you to create a wordlist from the results and then re-use it later in your searches.



Corpus of Historical American English



SEARCH

FREQUENCY

CONTEXT

ACCOUNT

SEE CONTEXT: CLICK ON WORD (ALL SECTIONS), NUMBER (ONE SECTION), OR [CONTEXT] (SELECT) [\[HELP...\]](#)

[iWeb](#)

[DISTURB](#)

	<input type="checkbox"/>	CONTEXT	ALL <input type="checkbox"/>	1810 <input type="checkbox"/>	1820 <input type="checkbox"/>	1830 <input type="checkbox"/>	1840 <input type="checkbox"/>	1850 <input type="checkbox"/>	1860 <input type="checkbox"/>	1870 <input type="checkbox"/>	1880 <input type="checkbox"/>	1890 <input type="checkbox"/>	1900 <input type="checkbox"/>	1910 <input type="checkbox"/>	1920 <input type="checkbox"/>	1930 <input type="checkbox"/>	1940 <input type="checkbox"/>	1950 <input type="checkbox"/>	1960 <input type="checkbox"/>	1970 <input type="checkbox"/>	1980 <input type="checkbox"/>	1990 <input type="checkbox"/>	2000 <input type="checkbox"/>
1	<input type="checkbox"/>	PEACE	199	2	6	22	17	12	11	15	10	21	13	15	14	4	8	7	6	3	3	7	3

0.719 seconds



SEARCH

ERROR

CONTEXT

ACCOUNT

SECTION: 1890 (21) (SHUFFLE)

CLICK FOR MORE CONTEXT



[?]

SAVE LIST

CHOOSE LIST



CREATE NEW LIST



[?]

SHOW DUPLICATES

1	1890	FIC	Cigarette-Makers	A	B	C	partner of her bliss returned -- presumably pacified by the soothing converse of his friend -- she would not disturb his peace of mind by any reference to the C
2	1890	MAG	NewEngMag	A	B	C	by more timid men, who dared go just far enough to commit themselves, disturb the peace of the state, and provoke the Law and Order government, but
3	1891	MAG	Atlantic	A	B	C	a government " having the confidence of the legislature. " Fiscal questions did not disturb the peace of parties in those days; the political battle Was fought on
4	1891	NF	Arena Volume4	A	B	C	at the stir which ideas are making in the community, and which threaten to disturb the peace and quiet of their mediocre godliness; and pious women engage
5	1892	FIC	TakenAlive	A	B	C	" It would be a pity to disturb your serenity. " " Nothing shall disturb it to-day. Peace is one of the rarest experiences in this world. I
6	1892	MAG	NorthAmRev	A	B	C	her Czars. She might be deterred from entering on her purpose lest she should disturb the peace of Europe. But India lies at Russia's very door with every
7	1892	NEWS	NYT-Reg	A	B	C	moan over and fondle involves many nice points; but it is not calculated to disturb the European peace. Oscar's threat, following George Moore's abortive den
8	1894	FIC	LostLadyLone	A	B	C	they would be perfectly safe. The next few days passed without anything occurring to disturb the peace of this misguided peasant girl. Every morning the man
9	1894	FIC	LostLadyLone	A	B	C	me to write and stop all my newspapers, which only brought me news to disturb my peace of mind. " I followed the direction of my wise guide.
10	1894	FIC	CasaBraccioVolumes122	A	B	C	. " It makes no difference, " answered the elder lady. " You disturb the peace of the sisters with your singing. You know the rule, and
11	1894	NEWS	NYT-Ed	A	B	C	he exacted at the close of the Franco-Prussian war, to make France powerless to disturb the peace of Germany for twenty years thereafter. That terni is now o
12	1894	NF	VoyageLiberdade	A	B	C	that part of the Brazilian coast fanned by constant trade-winds. Nothing unusual occurred to disturb our peace or daily course, and we pressed forward night
13	1895	FIC	Clarence	A	B	C	his wife, and others charged with inciting to riot and unlawful practice calculated to disturb the peace of the State of California and its relations with the Feder
14	1895	FIC	WarWithPontiac	A	B	C	expedition remained comfortably in camp, the weather was perfect, and nothing occurred to disturb the peace or enjoyment of the long voyage. Its only draw
15	1895	FIC	Mermaid ALove	A	B	C	him. The one day that he dare not listen long, that he must disturb her peace, was the only time that she had seemed to wish to make
16	1896	FIC	HolyCrossOther	A	B	C	tones, " you are sadly mistaken if you think you will be permitted to disturb our peace and harmony with your constant sighs and groans. If you are ill
17	1896	MAG	Century	A	B	C	and warning them of what they might expect under his rule if they continued to disturb the public peace. This was followed by energetic measures against the
18	1898	FIC	ArsRecteVivendi	A	B	C	dropped, and that he is Sardanapalus will not save him. If his revels disturb the college peace, he will be warned and dismissed. All that can be
19	1899	MAG	Atlantic	A	B	C	of change in the economy they administer; have a horror of everything that might disturb " the peace of the church, " belie*ng fervently in the maxim: "
20	1899	NEWS	NYT-Reg	A	B	C	is quiet along the entire line, nothing having happened up to this hour to disturb the peace of Sunday. In Manila the inhabitants have generally recovered from
21	1899	NF	HistoryEnglish	A	B	C	fruits, and whose continuance might in the present temper of Russia and its Czar disturb that peace of the Continent on which all his plans against England re

Corpus of Founding Era American English (COFEA)

Corpus Purpose:

This corpus is designed to represent general written American English from the founding era of the United States of America (i.e., 1765-1799). This corpus attempts to represent general writing by sampling language from multiple registers (see Biber, 1993). Biber (1993) argues that register diversity more so than corpus size is useful for general language studies because language can vary so vastly from one register to register. Therefore, register is a key variable that must be considered when designing interpreting results from corpora. Thus, although this corpus does not fully represent American English from the founding era because it is both large and register-diversified, it is currently the best corpus in existence for representing written language from that time period. We provide a detailed description of the composition of this corpus below.

Current Status:

Version 3.00 was built 4 February 2019.

It includes corrections of OCR errors and adjusted word counts.

Current sources include 119,801 texts from three sources for a total of 133,488,113 words.

Source	Documents	Words
Evans Early American Imprints	2,645	62,660,171
Founders Online	115,408	37,057,114
HeinOnline	277	32,237,273
Farrands	847	689,755
United States Statutes at Large	479	470,345



Constitution

<https://lcl.byu.edu/projects/cofea/>



State v. Rasabout, 356 P.3d 1258
(Utah 2015)
(Lee, J. Concurring)

"I write separately, however, because I cannot resolve the ambiguity in the term discharge as the court does—by mere resort to the dictionary. I see the need to look elsewhere. **I would interpret the terms of the statute by looking for real-world examples of its key words in actual written language in its native context. This sort of analysis has a fancy name—corpus linguistics.** But it is hardly unusual. We often resolve problems of ambiguity by thinking of examples of the use of a given word or phrase in a particular linguistic context. **I propose to do that (as I have in a couple of prior opinions) on a systematic scale—by computer-aided searches of online databases in an effort to assemble a greater number of examples than I can summon by memory on my own.**"



Wilson v. Safelite Grp., Inc., 930
F.3d 429, 438-39 (6th Cir. 2019)

“We ought to embrace another tool to ascertain the ordinary meaning of the words in a statute. This tool---corpus linguistics---draws on the common knowledge of the lay person by showing us the ordinary uses of words in our common language. How does it work? **Corpus linguistics allows lawyers to use a searchable database to find specific examples of how a word was used at any given time.** These databases, available mostly online, contain millions of examples of everyday word usage (taken from spoken words, works of fiction, magazines, newspapers, and academic works). Lawyers can search these databases for the ordinary meaning of statutory language like “results in.” The corresponding search results will yield a broader and more empirically-based understanding of the ordinary meaning of a word or phrase by giving us different situations in which the word or phrase was used across a wide variety of common usages. In short, corpus linguistics is a powerful tool for discerning how the public would have understood a statute’s text at the time it was enacted.”

State v. Lantis, 165 Idaho 427,
447 P.3d 875, 880 (2019)



Courts are beginning to utilize corpus linguistics as a means to aid the interpretation of statutory language in context and with the use of the empirical data available through extensive corpora, which catalogue millions of words. . . . One of the chief benefits of a corpus-linguistics-style analysis is that it offers a systematic, non-random look at the way words are used across a large body of sources.

We agree with these sentiments, but we recognize that the parties did not argue these principles in their briefing or at oral argument. We simply reference the use of corpus linguistic tools as a support for our analysis set forth above, and as an motivation for counsel to consider this “potential additional tool for our statutory interpretation toolbox,” [*State v. Rasabout*, 356 P.3d 1258 (Utah 2015) (Durham, C.J. Concurring),] when called for in the future.



The Utah Supreme Court encouraged lawyers to “provide courts with meaningful tools using the best available methods when the court is tasked with determining ordinary meaning”

Fire Ins. Exch. v. Oltmanns, 416 P.3d 1148, 1163 n.9 (Utah 2018).



*Wilson v. Safelite Grp.,
Inc.*, 930 F.3d 429, 438-
39 (6th Cir. 2019)

Of course, corpus
linguistics is one
tool---new to
lawyers and
continuing to
develop---but not
the whole toolbox.

Additional Resources

- Stephen C. Mouritsen, *Contract Interpretation with Corpus Linguistics*, 94 Wash. L. Rev. 1337 (2019)
- Thomas R. Lee & Stephen C. Mouritsen, *Judging Ordinary Meaning*, 127 YALE L.J. 788 (2018).
- Stephen C. Mouritsen, *Hard Cases and Hard Data: Assessing Corpus Linguistics as an Empirical Path to Plain Meaning*, 13 COLUM. SCI. & TECH. L. REV.156 (2011).
- Stephen C. Mouritsen, Note, *The Dictionary Is Not a Fortress: Definitional Fallacies and a Corpus-Based Approach to Plain Meaning*, 2010 BYU L. REV. 1915
- James Heilpern, Senior Fellow at BYU Law School:
heilpernj@law.byu.edu



QUESTIONS?